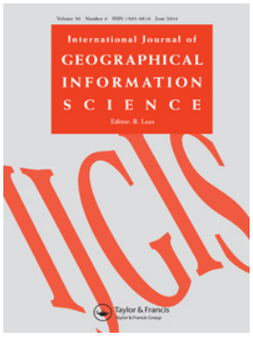

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Spatial scale and small area population statistics for England and Wales

Christopher D. Lloyd

To cite this article: Christopher D. Lloyd (2016) Spatial scale and small area population statistics for England and Wales, International Journal of Geographical Information Science, 30:6, 1187-1206, DOI: 10.1080/13658816.2015.1111377

To link to this article: http://dx.doi.org/10.1080/13658816.2015.1111377

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2016
VOL. 30, NO. 6, 1187–1206
http://dx.doi.org/10.1080/13658816.2015.1111377

Taylor & Francis
Taylor & Francis Group

Spatial scale and small area population statistics for England and Wales

Christopher D. Lloyd

Department of Geography and Planning, School of Environmental Sciences, University of Liverpool, Liverpool, UK

ABSTRACT

It is well-known that the results of analyses of aggregate data, such as those provided as outputs from censuses, are dependent on the size and shape of the zones used to report the data. However, many users of aggregate census data do not consider how far the zones utilised in their analyses capture spatial information about the population sub-groups they are studying. In addition, future data collection strategies should account for such issues. This article takes as its focus England and Wales, and it seeks to assess how far output areas (OAs) and aggregations of OAs capture information in selected population sub-groups and, therefore, how important it might be to use zones of a particular size in order to properly analyse the geographies of these sub-groups. The article uses the index of dissimilarity, D_{xy} (for groups x and y), and variograms to assess spatial variation in population sub-groups as represented by counts for OAs, lower layer super output areas (LSOAs), middle layer super output areas (MSOAs) and local authority districts (LAs), all produced as outputs from the Census in England and Wales. The analyses show how much information is contained at each spatial scale for sub-categories relating to age, ethnic group, housing tenure, car or van availability, qualifications, employment, limiting long-term illness (LLTI) and National Statistics Socio-economic Classification (NS-SeC). The amount of variation contained by each level of the hierarchy of zones differs markedly by population sub-group. LAs capture most (83%) of the variation in the spread of the population by the binary categorisation of White/Black, Asian and Minority Ethnic (far more than for any other variable), but there remains considerable local variation. The results suggest that zones larger than OAs are not geographically detailed enough to enable meaningful analysis of local-level differences between places and thus any alternative to the Census in the United Kingdom (with England and Wales as a specific case) must provide zones equivalent in size to OAs. If estimates are available only for larger areas then much information will be lost and our ability to explore how sub-group characteristics, or the relationships between variables, differ between localities will be considerably diminished. The results also provide evidence on some of the ways in which the population of England and Wales was geographically distributed in 2011.

ARTICLE HISTORY
Received 17 March 2015
Accepted 18 October 2015

KEYWORDS
Scale; spatial statistics; census data

CONTACT Christopher D. Lloyd c.d.lloyd@liverpool.ac.uk

1. Introduction

Recent debates about the future of the Census in the United Kingdom (see ONS 2013a) promoted analyses which sought to consider alternatives to a traditional census, as conducted in the United Kingdom in 2011. One key concern related to how far small area data derived from administrative sources, as opposed to a census, would provide sufficient *spatial* precision. For example, would estimates based on a sample provide usable small area counts? A short list of eight options for beyond 2011 was put forward by the Office for National Statistics (ONS) – these were grouped into options which were similar to the Census of 2011, administrative data options and survey options. These were subsequently reduced to two main options – (i) an online Census once a decade and (ii) combined use of administrative data and surveys (ONS 2013b). A key advantage of the latter would be the ability to chart population change over short time periods in contrast with a decadal census model. Assessments of the latter were based on data collected by bodies including the Department of Health, the Department for Work and Pensions, HM Revenue and Customs, the Department for Education, the Higher Education Statistics Agency, NHS Wales and the Welsh Government. The proposed approach would have included an annual survey of some 1% of households to adjust for those not included in the administrative data sources or recorded in the incorrect location, with a second annual survey of around 4% of households capturing information on characteristics not captured in the administrative data sources – these include ethnicity and languages spoken. This approach would allow for annual population estimates at local authority level and, after the survey has run for several years, estimates for smaller areas such as wards. While this approach would lead to estimates by age and sex for small groups of postcodes, the array of detailed OA-level statistics as provided via the Census would not be available. The two options were assessed in greater depth in ONS (2014) in the context of a public consultation about the options. Similar questions have arisen in the United States, where the traditional long form Census has been replaced with a short-form Census and the American Community Survey (ACS) and estimated counts, along with a margin of error, are provided for small areas (see Spielman *et al*. 2014 for an assessment of uncertainty in ACS data). This article is concerned with the spatial information loss associated with cases such as combined use of administrative data and surveys as outlined by the ONS (2013b). The article takes as its specific focus geographical variation in population sub-groups in England and Wales and it seeks to determine how much spatial information would be lost if output area (OA) level data (released as outputs from the 2011 UK Census) were not available and instead spatial aggregations of OAs were the finest geographies at which counts were released. It is worth stressing that the article takes the view that there is no prior reason to prefer a census over administrative data-based linkages if the two provide equivalent spatial detail and precision of measurement for attributes of interest.

A lack of sufficient spatial detail creates problems for any application which is reliant on detailed spatial information (e.g. targeting area deprivation, analyses of socio-economic or ethnic segregation or spatial regression modelling). More generally, this means that we may lack sufficient information on the changing geographies of some population sub-groups and thus lack vital details about an important facet of social or economic change. Taking an example, if we wish to explore the relationship between

deprivation and ethnicity then we require reliable individual or small area data on both characteristics in order to properly assess this relationship and insufficient spatial detail may render the analysis meaningless. Resource allocation by government depends on knowledge of the scales over which population sub-groups (e.g. those with poor health) are distributed. As an example, NHS England allocates funds to Clinical Commissioning Groups (CCGs, of which were are 211 in England as of March 2013[1]) based on the size of the population of each CCG with weights per head of the population derived based on need due to age, additional need based on health status, an adjustment for unmet need and health inequalities and also unavoidable costs associated with healthcare delivery associated with location alone (e.g. staff, land and building costs being higher in London than elsewhere). NHS England (2014) states that the share of the overall weighted capitation formula accounted for by the unmet need adjustment is set at 10%. The unmet need adjustment is based on the standardised mortality ratio for people aged under 75 (SMR < 75), derived for middle layer super output areas (MSOAs) which are then aggregated to CCGs. The rationale behind this is that inequalities *within* as well as *between* CCGs are then taken into account. A key problem with such an approach is that there is no rational basis behind the selection of MSOAs and if SMR < 75 varies markedly within MSOAs, then there may be a considerable mismatch between unmet need and the amount allocated in practice. This connects to the notion of the ecological fallacy – the fallacy of making inferences about individuals from aggregate data (Lloyd 2014).

The modifiable areal unit problem (MAUP) lies at the heart of any analysis of aggregate data (such as population counts for census areas) – the results of any analysis are, in part, a function of the support (geometrical size, shape and orientation of the measurement units; Atkinson and Tate 2000) over which the spatial aggregation takes place and the article connects to previous research about the MAUP, population surveys and geographical patterns in population sub-groups in England and Wales. Openshaw and Taylor (1979) and Openshaw (1984) consider the impact of choice of spatial aggregation on univariate and multivariate statistical analyses. Wong (2009), Lloyd (2014) and Manley (2014) provide introductions to the MAUP while Fotheringham *et al*. (2002) consider its implications for local regression analysis. The particular focus in this article is on characterising how population sub-groups are spatially distributed – for example, how large, on average, are differences within regions as opposed to differences between areas? The article builds on research conducted by, among others, Voas and Williamson (2000), Dorling and Rees (2003) and Lloyd (2015). Voas and Williamson (2000) sought to assess unevenness in population groups using 1991 Census data for England and Wales. Using data for districts, wards and enumeration districts, they considered what proportion of unevenness in the selected paired population sub-groups was found at each of the three scales represented by these zones. Dorling and Rees (2003) explored changes in spatial divisions across Britain between 1971 and 2001. In the study by Lloyd (2015), the spatial structure of population sub-groups in England and Wales in 2001 and 2011 was the focus and the analyses suggested that differences between regions, in terms of most population sub-groups considered, had reduced between 2001 and 2011. Other studies have focused on changes by ethnicity (e.g. Catney 2015) and demographic and deprivation change (Norman 2010). The present study builds on this work by assessing the possible implications of choosing alternative zones and by assessing systematically the scales over which

population sub-groups are distributed across England and Wales. The research has direct links to analyses of residential segregation whereby the aim is to assess how far members of different population sub-groups tend to live either together or apart (see Lloyd *et al*. 2014 for a recent overview).

Previous work on scale effects and population variables has included research on segregation measured at multiple spatial scales (Wong 1997, Voas and Williamson 2000, Reardon *et al*. 2008, 2009) and the work of Tranmer and Steel (2001) and Manley *et al*. (2006), which uses a multilevel modelling framework to assess scale effects in Census variables. Grasland *et al*. (2000) consider scale effects in the representation and analysis of populations while Griffith *et al*. (2003), in an analysis of population density in the United States, explore the relationship between different measures of spatial autocorrelation for different zonal systems and between global and local measures. There are many reviews of the components of the MAUP which use population data to illustrate key issues or as motivating examples (recent examples include Lloyd 2014 and Manley 2014). Others have attempted to reallocate census counts from one zonal system to another or from a zonal system to a grid (e.g. Martin 1989), thus freeing analysts from the constraints of the geographical zones used to report counts (but, of course, with the limitation that the end results are estimates). The present study builds on previous research by making use of different methods (including variograms) to illustrate how variables are spatially structured at different scales and by making use of the most recent (2011) Census data for England and Wales as made available for four different sets of zones.

A key issue which motivates this research relates to determining the smallest set of zones which could be used to represent local area geographical variation in population sub-groups in England and Wales. As an example, if LSOAs are used in place of OAs in an analysis of the geography of housing tenure, is the loss of spatial information such that the analysis may markedly under-represent geographical variation by housing tenure relative to an analysis based on OAs? One approach is to determine internally homogeneous regions (e.g. Folch and Spielman 2014) which capture variation in the property of interest. Much previous work on scale change refers to images and focuses on optimal spatial resolutions (where cells in a coarser resolution image each contain the same number of cells at a finer resolution). In the present case, the number of OAs in a LSOA (for example) is not constant and neither is the size of the zones. Atkinson and Curran (1997), who are concerned with remotely sensed imagery, view the optimal spatial resolution as one which captures spatial variation of interest but where there is little redundancy (in the latter case, neighbouring pixels tend to be similar and the local (moving window) variance is small). Where identifying an appropriate scale of measurement is the aim and data are not available on a point support (i.e. in the present case, by household), the most suitable approach is to estimate the point support variogram from the areal variogram (e.g. the variogram estimated from OA-level data) and to use a deconvolution approach to estimate the point support variogram. The point support variogram can then be regularised to determine how the variogram changes as the spatial resolution decreases (cells become larger) and an optimal cell size could be identified in this way. Variogram deconvolution is possible where the data are for irregularly shaped zones rather than cells (see Goovaerts 2008 and also Lloyd 2014, Zhang *et al*. 2014). However, in the present study, the aim is not to identify an optimal

size (and shape) of zones but to assess how much information is contained within each set of zones assessed (the index of dissimilarity and geographical variances are computed to determine how much variation is associated with each scale, as represented by the nested zonal systems). Therefore, the variogram (and other measures) are derived using a set of zones for which Census counts are provided. OAs are taken as a starting point, as they are the smallest areas for which Census counts are released in the United Kingdom. The method used for construction of OAs is summarised in Section 2.

The analysis is based on counts released for output areas (OAs; n = 181,408; mean population = 309), lower layer super output areas (LSOAs; n = 34,753; mean population = 1614), middle layer super output areas (MSOAs; n = 7201; mean population = 7787) and local authority districts (LAs; n = 348; mean population = 161,138; but note that LAs are used only in selected parts of the analyses). The variables used are derived from counts by age, ethnic group, housing tenure, car or van availability, qualifications, employment, limiting long-term illness (LLTI) and National Statistics Socio-economic Classification (NS-SeC). The analyses presented in this article comprise two main parts:

(1) Characterising unevenness using the index of dissimilarity, D_{xy}(for groups x and y)
(2) Measurement of spatial structure using variograms estimated from log-ratios (derived from percentages of people in particular sub-groups)

First, the data used in the analysis are described. Next, the analyses are summarised in three sections – geographical unevenness, computing log-ratios and analysing spatial variation with the variogram. The analysis seeks to demonstrate how crucial small areas are if we are to capture spatial variation in population sub-groups. The analyses also provide a rich picture of the geographical distribution of the selected population sub-groups in England and Wales.

2. Data

As noted above, the analysis makes use of counts for OAs, LSOAs, MSOAs and LA districts for England and Wales. While the analyses could, in principle, be extended to include the rest of the United Kingdom, the decision was taken to focus on England and Wales given difficulties associated with analysing datasets with characteristics which vary between the countries of the United Kingdom (e.g. data zones in Scotland rather than LSOAs). Future work should consider these differences and seek to assess scale effects across the whole of the United Kingdom. LAs are used only in the first part of the analysis (assessment of geographic unevenness) as, for the second part of the analysis (using variograms), LAs were considered far too large to provide a meaningful comparison with the three other sets of zones. OAs were first used in the 2001 Census in England and Wales and they were constructed using clusters of adjacent unit postcodes; they were intended to have similar population sizes and to be as socially homogenous as possible according to tenure of household and dwelling type.[2] The automated zone design methodology used to generate OAs is detailed by Martin *et al.* (2001). Super OAs are built from groups of OAs and they were developed so as to enable government departments to report statistics at a relatively fine spatial scale but with a small risk of disclosure of information about individuals.[3] The data used in the current analysis, as

Table 1. Key statistics census tables and derived variables.

Table	Table description	Description
KS102	Age structure	Age 0 to 15; 16 to 29; 30 to 64; 65 plus
KS201	Ethnic group	Whites; Black, Asian and Minority Ethnic
KS402	Housing tenure	Owner occupied; social rented; private rented
KS404	Cars and vans	Cars or vans; no cars or vans
KS501	Qualifications and students (persons aged 16 and over)	No qualifications; qualifications
KS601	Economic activity – all persons (aged 16–74)	Unemployed economically active; employed economically active
KS301	Health and provision of unpaid care	With LLTI; no LLTI
KS611	NS-SeC (persons aged 16–74)*	NS-SeC1, 2; 3 to 7; 8

Note: *The NS-SeC classes are as follows:
NS-SeC 1, 2: Managerial, administrative and professional occupations
NS-SeC 3–7: Intermediate, routine and manual occupations
NS-SeC 8: Never worked and long-term unemployed.

specified in Table 1, are from the Key Statistics tables. The variables are selected as they encompass important population groupings and they represent a diverse array of spatial characteristics, with, for example, some variables being strongly spatially clustered while others have only weak spatial structure (i.e. there are no obvious spatial trends such as distinct concentrations of a sub-group in urban as against rural areas). The groupings used (e.g. age ranges) and number of variables are limited since the analysis is intended to be illustrative. Table 2 summarises counts and percentages for all of the variables used in this analysis. Lloyd (2015) used the same variables and summarised changes in the same population sub-groups between 2001 and 2011.[4] Population-weighted centroids (British National Grid coordinates in metres) are used in the variogram analyses presented below. It should be noted that centroids may be a poor approximation of what should be a continuous population surface, especially in the case of larger zones. Alternative approaches which seek to overcome such limitations are offered by Mockus

Table 2. Counts and percentages.

Variable	Definition	2011	2011%
A0to15	Persons aged 0 to 15	10,579,132	18.87
A16to29	Persons aged 16 to 29	10,495,245	18.72
A30to64	Persons aged 30 to 64	25,778,462	45.97
A65plus	Persons aged 65 plus	9,223,073	16.45
White	White persons	48,209,395	85.97
BAME	Black, Asian and Minority Ethnic	7,866,517	14.03
OwnOcc	Owner occupied HH	15,031,914	64.33
SocRent	Social rented HH	4,118,461	17.63
PrivRent	Private rented HH	4,215,669	18.04
CarsVans	HH with cars or vans	17,376,274	74.37
NoCarsVans	HH with no cars or vans	5,989,770	25.63
Qual	Persons with qualifications	35,189,453	77.34
NoQual	Persons with no qualifications	10,307,327	22.66
EAEmploy	EA employed persons	25,449,863	93.40
EAUnemp	EA unemployed persons	1,799,536	6.60
NSSEC12	NS-SeC 1,2	12,792,224	34.19
NSSEC37	NS-SeC 3–7	22,324,839	59.66
NSSEC8	NS-SeC 8	2,301,614	6.15
NoLLTI	Persons with no LLTI	46,027,471	82.08
WithLLTI	Persons with a LLTI	10,048,441	17.92

Note: HH are households; EA is economically active. NB: PrivRent includes 'Private rented: Private landlord or letting agency', 'Private rented: Other' and 'Living rent free'; NS-SeC counts include imputed persons.

(1998), Goovaerts (2008) and Nagle *et al*. (2011). However, population-weighted centroids of OAs, LSOAs and MSOAs were accepted in this analysis since, for these zones, random reassignment of centroids within zones would make minimal difference to the results.

3. Analysis

The analysis is divided into two parts. The first uses raw counts of persons by group to measure unevenness in the population sub-groups (supported by an analysis of geographical variances). The second part is based on log-ratio transformations of percentages of people in an area who belong to a given group. Percentages are computed with $x_i/t_i \times 100$, where x_i is the number of persons in a sub-group in area i and t_i is the total number of persons in area i. Figure 1 provides, for context, maps of the percentage of Black, Asian and Minority Ethnic (BAME) persons (Figure 1A) and the percentage of persons with no qualifications (Figure 1B) by OA. It is clear that the percentages of BAME persons tend to be larger in urban areas while the spatial patterns observable in the map of percentages of persons with no qualifications are rather different with no obvious urban/rural distinction. In short, the spatial structures of the two variables are different and the following analyses seek to summarise how the population is spatially distributed by an array of characteristics.

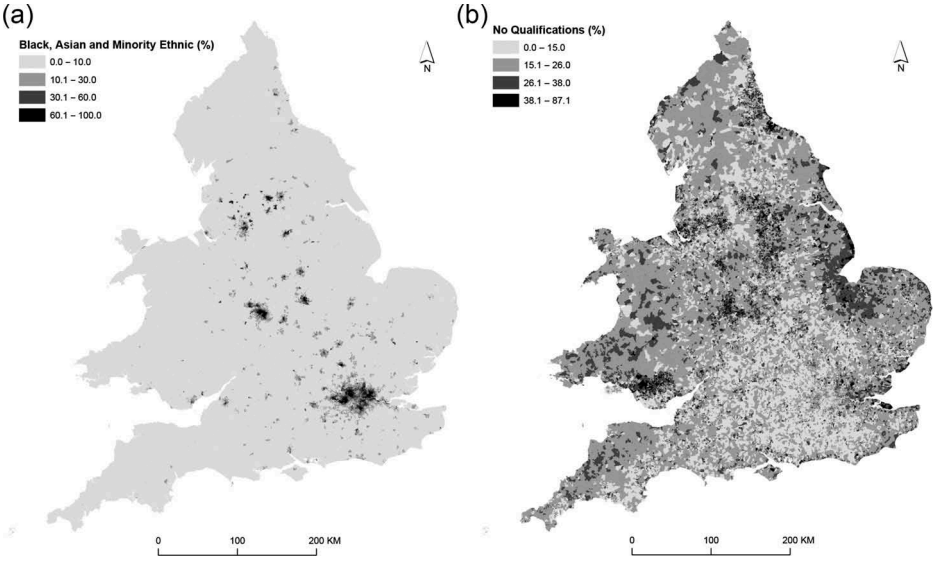

Figure 1. (A) Percentage of Black, Asian and Minority Ethnic persons by OAs. (B) Percentage of persons with no qualifications by OAs. Contains National Statistics data © Crown copyright and database right 2012. Contains Ordnance Survey data © Crown copyright and database right 2012.

3.1. *Geographical unevenness*

The index of dissimilarity, D_{xy}, summarises the total differences between the spread of the population sub-groups x and y over all of the areal units:

$$D_{xy} = 0.5 \times \sum_{i=1}^{n} \left| \frac{x_i}{X} - \frac{y_i}{Y} \right| \quad (1)$$

where x_i and y_i are counts of the population in two sub-groups for areal unit i and there are n units. X and Y are the total population counts of each sub-group across the whole of the study area. D_{xy} takes a value between 0 and 1 where a small value indicates evenness and a large value suggests a high degree of unevenness (see Duncan and Duncan 1955, Massey and Denton 1988).

Voas and Williamson (2000) compute the percentage contributed by each of several geographical levels and their approach is used here (Table 3; abbreviations are defined in Table 2). Note that the zones used in this study represent a nested hierarchy as is required for this approach to be meaningful. The contributions are computed as follows:

$$\begin{aligned} D_{xy}(\text{LA})\% &= 100 \times (D_{xy}(\text{LA})/D_{xy}(\text{OA})); \\ D_{xy}(\text{MSOA})\% &= 100 \times ((D_{xy}(\text{MSOA}) - D_{xy}(\text{LA}))/D_{xy}(\text{OA})); \\ D_{xy}(\text{LSOA})\% &= 100 \times ((D_{xy}(\text{LSOA}) - D_{xy}(\text{MSOA}))/D_{xy}(\text{OA})); \\ D_{xy}(\text{OA})\% &= 100 \times ((D_{xy}(\text{OA}) - D_{xy}(\text{LSOA}))/D_{xy}(\text{OA})) \end{aligned}$$

Only for White/BAME is less than 10% of the variation contributed by OAs. The results suggest that most of the variation in White/BAME is captured by LAs (over 83%). For A0to15, A30to64 and A65plus (as against other ages), SocRent (as against other housing tenures) and LLTI, over 20% of the variation is contributed by OAs. These results suggest that all of the four sets of zones could be used to assess the overall distribution of the population using the binary classification of White/BAME. Of course, there are considerable local variations, and also the geographies of individual ethnic groups are far more complex than the simple two-part grouping used here. In the case of (No)CarsVans and

Table 3. Index of dissimilarity, *D* (for stated sub-group versus remainder in category (e.g., A0to15 versus all others by age and White versus BAME)).

	D				% contributed by level			
	OA	LSOA	MSOA	LA	LA	MSOA	LSOA	OA
A0to15	0.161	0.119	0.089	0.048	29.814	25.466	18.634	26.087
A16to29	0.208	0.178	0.159	0.115	55.288	21.154	9.135	14.423
A30to64	0.102	0.075	0.059	0.035	34.314	23.529	15.686	26.471
A65plus	0.274	0.209	0.167	0.114	41.606	19.343	15.328	23.723
White	0.592	0.576	0.564	0.494	83.446	11.824	2.027	2.703
OwnOcc	0.446	0.370	0.303	0.178	39.910	28.027	15.022	17.040
SocRent	0.592	0.468	0.355	0.187	31.588	28.378	19.088	20.946
PrivRent	0.371	0.308	0.260	0.153	41.240	28.841	12.938	16.981
NoCarsVans	0.402	0.357	0.317	0.229	56.965	21.891	9.950	11.194
NoQual	0.255	0.212	0.181	0.117	45.882	25.098	12.157	16.863
EAUnEmp	0.300	0.252	0.219	0.145	48.333	24.667	11.000	16.000
NSSEC12	0.265	0.237	0.213	0.134	50.566	29.811	9.057	10.566
NSSEC37	0.207	0.185	0.167	0.114	55.072	25.604	8.696	10.628
NSSEC8	0.374	0.329	0.296	0.216	57.754	21.390	8.824	12.032
WithLLTI	0.199	0.150	0.122	0.093	46.734	14.573	14.070	24.623

NSSEC, a clear majority of the variation is at LA and MSOA levels. For sub-groups by Age and Tenure, and for LLTI, >20% of the variation would be effectively discarded if zones larger than OAs were used. In these cases in particular, the proper analysis of, for example, geographical inequalities would be strongly dependent on the use of OAs rather than any larger spatial aggregation of counts.

The results using D_{xy} are supported by an analysis based on the geographical variances approach of Moellering and Tobler (1972, and see also Silk 1981). This approach is based on the idea that the sums of squares of deviations of population counts at each geographical level sums to the total sums of squares (the variation around the grand mean) and the total variability can be divided by the sum of squares at each level. An analysis based on all of the sets of counts used in this article suggests that most of the 'action' for most counts takes place at the OA level. Only for the BAME population is there is there strong evidence that most action is found at a coarser level – for LAs in this case[5] (the full results are not presented due to space limitations).

3.2. *Computing log-ratios*

The second part of the analysis is based on variograms computed from log-ratio transformed percentages. Percentages are constrained to sum to 100 (while proportions sum to one), and such data are referred to as compositional. Many studies have argued that statistical analysis of raw percentages or proportions is not appropriate, and transforming percentages or proportions to log-ratios provides one solution (see Lloyd *et al*. 2012 for an introduction to the topic in a population studies context). Several alternative log-ratio transforms of compositional data have been developed; the additive-log-ratio (alr) and the centred-log-ratio (clr) were defined by Aitchison (1986), but alr and clr transformed data are subject to restrictions in their treatment by standard methods. Isometric-log-ratio (ilr) transformed data (Egozcue *et al*. 2003 and see Egozcue and Pawlowsky-Glahn 2006, for an introduction) can be analysed directly using standard univariate or multivariate statistical methods and this approach is used in the analysis presented here. Balances are a particular form of ilr coordinates (Egozcue and Pawlowsky-Glahn 2005) and they represent the relative variation in two groups of parts. The present analysis is based on balances with parts (sub-groups expressed as percentages) defined as follows (with the number of parts indicated):

Two part compositions:

$$\text{Ethnicity} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \ln\frac{\text{White}}{\text{BAME}},\ \text{CarsVans} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \ln\frac{\text{NoCarsVans}}{\text{CarsVans}},$$

$$\text{Qual} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \ln\frac{\text{NoQual}}{\text{Qual}},\ \text{Employ} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \ln\frac{\text{EAEmploy}}{\text{EAUnemp}},$$

$$\text{LLTI} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \ln\frac{\text{WithLLTI}}{\text{NoLLTI}}$$

Three part compositions:

$$\text{Tenure 1} = \left(\frac{2}{3}\right)^{\frac{1}{2}} \ln \frac{(\text{OwnOcc} \times \text{PrivRent})^{\frac{1}{2}}}{\text{SocRent}}, \quad \text{Tenure 2} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \ln \frac{\text{OwnOcc}}{\text{PrivRent}},$$

$$\text{NSSEC1} = \left(\frac{2}{3}\right)^{\frac{1}{2}} \ln \frac{(\text{NSSEC12} \times \text{NSSEC37})^{\frac{1}{2}}}{\text{NSSEC8}}, \quad \text{NSSEC2} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \ln \frac{\text{NSSEC12}}{\text{NSSEC37}}$$

Four part compositions:

$$\text{Age1} = \left(\frac{3}{4}\right)^{\frac{1}{2}} \ln \frac{(\text{A0to15} \times \text{A16to29} \times \text{A30to64})^{\frac{1}{3}}}{\text{A65plus}},$$

$$\text{Age2} = \left(\frac{2}{3}\right)^{\frac{1}{2}} \ln \frac{(\text{A0to15} \times \text{A16to29})^{\frac{1}{2}}}{\text{A30to64}}, \quad \text{Age3} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \ln \frac{\text{A0to15}}{\text{A16to29}}$$

Some counts are zeroes and, following Lloyd (2015), the percentages were calculated from counts x_1, x_2, x_3… with $x_1 + 1$, $x_2 + 1$, $x_3 + 1$… So, a value of one is added to all counts and the percentages, y_1, y_2, y_3… are calculated from the modified counts. Lloyd (2015) assessed the sensitivity of results to the addition of different values (e.g. 0.1 and 0.5) and the results were found to be robust. Note that MSOAs have no zero counts for any categories and so log-ratios (z_1, z_2, z_3…) were also computed for MSOAs without an addition of 1 to each count. In this case, the standard deviations of the log-ratios are the same, to two decimal places, as those based on counts with 1 added and thus the addition of 1 was judged to have little impact in this case.

3.3. *Analysing spatial variation with the variogram*

The variogram is used here to consider how the population sub-groups are spatially structured over multiple scales, but it also provides a rich perspective on how the choice of zones could affect analyses. The variogram, $\gamma(\mathbf{h})$, relates half the average of the squared differences between zones to the distances (in lags or bins) separating their centroids (here, population-weighted centroids). The variogram provides a summary of spatial dependence at different spatial scales. The experimental variogram can be estimated for the $p(\mathbf{h})$ paired observations (log-ratios in the present study), $z(\mathbf{s}_i)$, $z(\mathbf{s}_i + \mathbf{h})$, i= 1, 2, …, $p(\mathbf{h})$ with:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2p(\mathbf{h})} \sum_{i=1}^{p(\mathbf{h})} \{z(\mathbf{s}_i) - z(\mathbf{s}_i + \mathbf{h})\}^2 \tag{2}$$

where $z(\mathbf{s}_i)$ is an observation at the ith location $\mathbf{s}_i$ and $\mathbf{h}$ is the lag (distance and direction) by which two observations are separated. The variogram, like the other measures employed in this article, is a function of the data support (see Lloyd 2014), and, in the present analysis, it is computed from log-ratios derived from data for OAs, LSOAs and MSOAs, thus providing a means of assessing information contained at different spatial scales. The analysis of compositional data using variograms is discussed by Pawlowsky and Burger (1992) and Pawlowsky-Glahn and Olea (2004). The weighting of variograms by population size has been assessed by some researchers (see Goovaerts *et al*. 2005) and was tested by Lloyd (2015); in common with the latter, only unweighted variograms are used here.

Models are often fitted to variograms (particularly as an input to kriging interpolation; Webster and Oliver 2007). Fitted models also provide a useful summary of the structure of variograms and models are fitted to the variograms estimated from OA, LSOA and MSOA data. The models fitted in this analysis comprise two elements – a nugget effect and one or two spherical model components (but some fitted models do not include a nugget effect). A spherical model component is defined by the range (denoted by a, representing the spatial scale of variation) and the structured component (c, representing spatially correlated variation); more than one spherical model component is fitted to some variograms which have more complex forms. The nugget effect, c_0, represents measurement error and variation at a distance smaller than that represented by the sample spacing (Lloyd 2012, 2014, discusses variogram estimation and modelling). The nugget effect plus the structure component(s) are the total sill (the *a priori* variance). The range indicates the spatial *scale* of variation while the nugget effect and structured component(s) indicate the *magnitude* of variation. Example models are described below with reference to the nugget effects, ranges and structured components (see Lloyd 2014 for a graphical illustration of a nugget effect and a spherical model component). The variograms were estimated using the Gstat software (Pebesma and Wesseling 1998). The models were fitted in Gstat using weighted least squares whereby the weights are a function of the number of paired observations at each lag.

Variograms add to the analyses using D_{xy} by taking into account spatial location (D_{xy} as applied here treats zones as isolated entities with, in effect, no interaction across boundaries) and by allowing the exploration of spatial concentrations of population sub-groups over multiple scales. Variograms for each variable (ilr transformed percentages) for OAs, LSOAs and MSOAs are shown in Figure 2 with models fitted in each case (in some cases the model does not extend to the maximum lag – this indicates that it was only fitted to some subset of the semivariances). Table 4 summarises the nugget effects (note that two models do not include nugget effects), structured components and the ranges. As an example, the OA-level variogram for ethnicity has a nugget effect (c_0) of 0.172; for the first spherical component it has a structured component (c_1) of 0.35 and a range (a_1) of 6263 m while, for the second spherical component, it has a structured component (c_2) of 0.69 and a range (a_2) of 56304 m. The range values indicate dominant spatial features at a local scale (the first range; approximately 6 km) and over a larger 'regional' scale (with a range of approximately 56 km). Smaller range figures would indicate more localised patterns while larger range figures suggest more 'gradual' spatial trends. Thus, variables which have short-range spatial variation (e.g. Age2, Age3, Qual and LLTI, for which there is little spatial structure over larger areas) can be distinguished from those which have much longer range spatial variation (e.g. Ethnicity and NSSEC1).

The variograms for OAs have larger maximum semivariances than those for the LSOA and MSOA data, while the maximum semivariances for the LSOA data are larger than those for the MSOA data. This indicates that the LSOAs smooth variation relative to the OAs, and, with MSOAs, there is a further loss of information. Larger differences in the maximum semivariance values indicate a larger degree of information loss with spatial aggregation from OAs to LSOAs and from LSOAs to MSOAs. The smallest proportional difference is for Ethnicity – the variograms for OAs, LSOAs and MSOAs are very similar in form and there are only quite

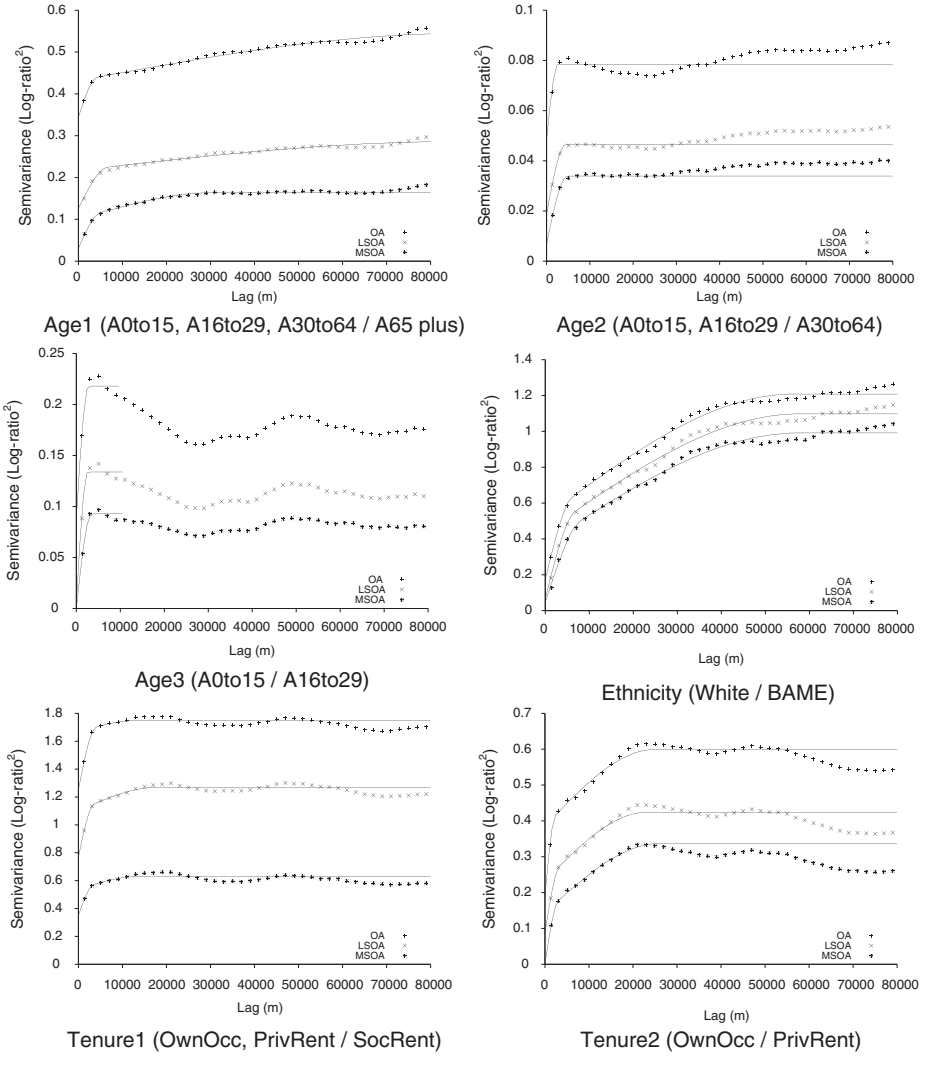

Figure 2. Variograms for all log-ratios for OAs, LSOAs and MSOAs (2 km lag).

small differences in maximum semivariances in this case. For several other variables, there are notably larger differences between semivariances for OAs and LSOAs than for LSOAs and MSOAs. In other words, more information is lost by aggregation from OAs to LSOAs than from LSOAs to MSOAs. The forms of the variograms (the relative values of the semivariances) for each log-ratio are similar in most cases for OAs, LSOAs and MSOAs. The rate of change in semivariances with increasing distance indicates the scale of variation (as captured by the range) – if semivariances increase markedly with an increase in distance then the variation is short range whereas a gradual increase in semivariance with distance suggests long-range spatial variation (corresponding to visually smooth mapped values with little difference over small distances). A large difference between semivariances at small lags and semivariances at large lags is suggestive of strong spatial structure (more obvious spatial patterns such as distinct urban–rural differences in sub-group populations) while a relatively 'flat' variogram,

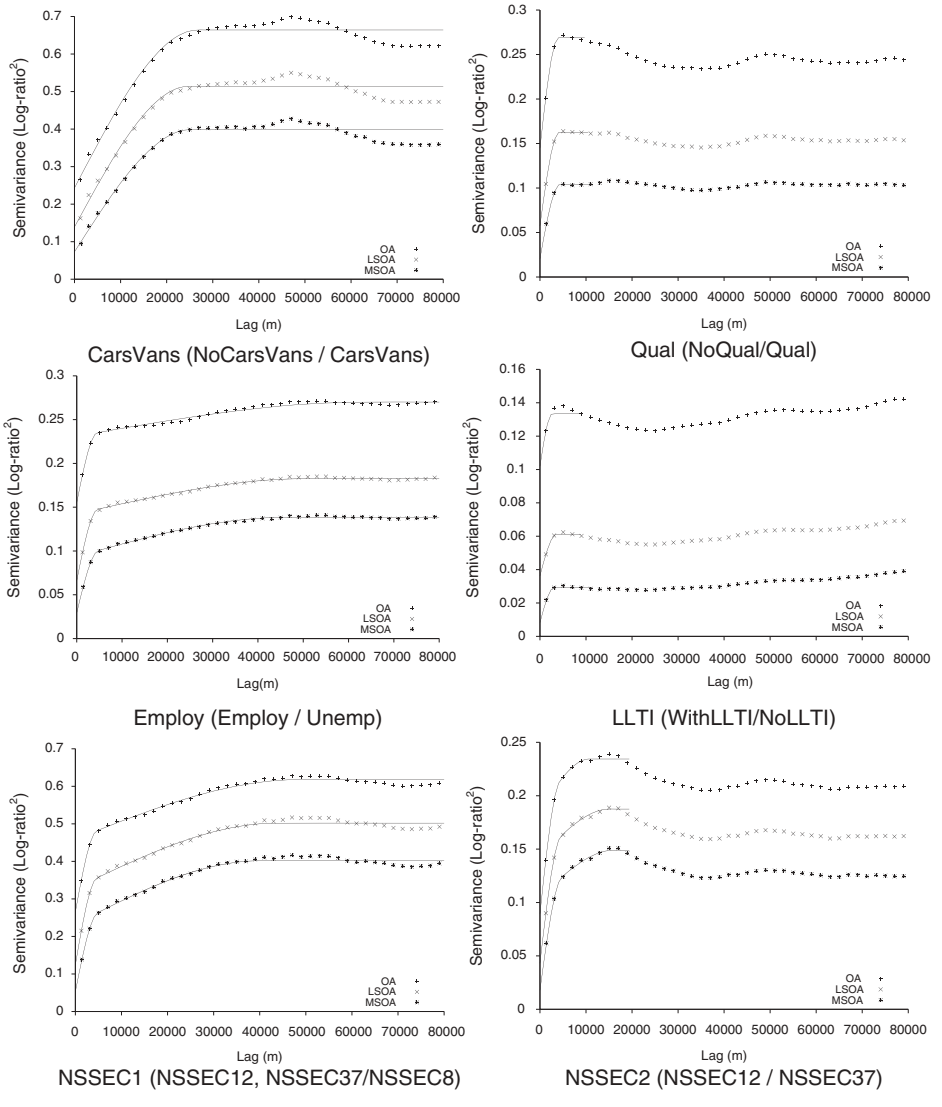

Figure 2. Continued.

with little difference between the semivariances at the smallest and largest lags indicates little spatial structure (less obvious spatial patterns). Thus, the variograms for Ethnicity display most evidence for strong spatial structure while those for LLTI indicate very little spatial structure. In summary, the major trends are captured by OAs, LSOAs and MSOAs, but the magnitude of spatial variation differs substantially across the three sets of zones for some variables. This suggests that LSOAs and MSOAs represent regional trends quite well for all of the broad groups included in this study but that the impact of aggregation differs between variables in terms of changes in variation (also shown by D_{xy}). It is no surprise that LSOAs and MSOAs tend to capture regional trends illustrated using OAs as the average size of the zones is small relative to the scale of analysis.

Table 4. Variogram model coefficients; 'sph' is spherical.

Zones	Variable	c_0	c_1	Model	a_1	c_2	Model	a_2
OA	Age1	0.344	0.087	sph	4263.96	0.113	sph	86839.5
LSOA	(A0to15, A16to29,	0.126	0.091	sph	6757.89	0.069	sph	85467.8
MSOA	A30to64/A65plus)	0.03	0.07	sph	4952.33	0.064	sph	31068.3
OA	Age2	0.048	0.03	sph	2621.18			
LSOA	(A0to15, A16to29/	0.019	0.028	sph	4477.59			
MSOA	A30to64)	0.007	0.027	sph	4836.4			
OA	Age3	0.089	0.129	sph	2773.67			
LSOA	(A0to15/A16to29)	0.005	0.129	sph	2560			
MSOA			0.093	sph	3438.41			
OA	Ethnicity	0.172	0.35	sph	6262.94	0.686	sph	56303.7
LSOA	(White/BAME)	0.066	0.373	sph	6846.92	0.659	sph	57609.8
MSOA		0.045	0.329	sph	8235.68	0.617	sph	57531.5
OA	Tenure1	1.251	0.412	sph	4014.59	0.087	sph	12878.5
LSOA	(OwnOcc, PrivRent/SocRent)	0.768	0.333	sph	3565.81	0.169	sph	18006.5
MSOA		0.349	0.189	sph	3472.66	0.094	sph	14197.3
OA	Tenure2	0.189	0.196	sph	2369.04	0.214	sph	24782.4
LSOA	(OwnOcc/	0.088	0.145	sph	3344.24	0.191	sph	22850.4
MSOA	PrivRent)		0.144	sph	3050.95	0.193	sph	26003.2
OA	CarsVans	0.243	0.421	sph	26846.5			
LSOA	(NoCarsVans/CarsVans)	0.139	0.375	sph	24757.4			
MSOA		0.073	0.326	sph	25275.1			
OA	Qual	0.145	0.124	sph	4011.36			
LSOA	(NoQual/Qual)	0.058	0.105	sph	4174.95			
MSOA		0.02	0.083	sph	4326.28			
OA	Employ	0.155	0.076	sph	4538	0.039	sph	63928.9
LSOA	(Employ/Unemp)	0.066	0.076	sph	4654.39	0.041	sph	51802.1
MSOA		0.029	0.064	sph	4784.13	0.045	sph	45947.6
OA	NSSEC1	0.265	0.194	sph	4645.67	0.159	sph	48765.6
LSOA	(NSSEC12, NSSEC37/NSSEC8)	0.124	0.204	sph	4644.74	0.174	sph	43760
MSOA		0.053	0.18	sph	4851.75	0.169	sph	40091.5
OA	NSSEC2	0.089	0.096	sph	4242.29	0.049	sph	10735.9
LSOA	(NSSEC12/	0.043	0.099	sph	4559.41	0.045	sph	15133.3
MSOA	NSSEC37)	0.018	0.087	sph	4701.14	0.044	sph	16135.2
OA	LLTI	0.102	0.031	sph	2621.18			
LSOA	(WithLLTI/NoLLTI)	0.036	0.025	sph	3573.12			
MSOA		0.008	0.021	sph	3048.61			

Note: The parts (percentages) given in parenthesis after the variables (log-ratios) are defined following Section 3.2; the numerators are separated by commas.

4. Discussion and conclusions

This article makes two key contributions. The first relates to spatial variation in population sub-groups (demographic and socio-economic) in England and Wales. The second concerns the amount of spatial information contained at different spatial scales for the (ilr-transformed) variables included in the analysis. With respect to the first contribution, the results support the work of Lloyd (2015; based on analysis of OA-level data) in showing how the spatial structure of population sub-groups in England and Wales varies. The ranking of D_{xy} values presented by Voas and Williamson (2000) is similar to that presented here in that unevenness by ethnicity (White/BAME) is largest while the population tends to be more evenly distributed by age. The article adds to previous work by showing how much variation is lost by moving from OAs to LSOAs or MSOAs and it is argued that OAs are required for meaningful analyses of most sub-groups. Only in the case of the White/BAME populations could it reasonably be argued that zones

larger than OAs might be satisfactory but even in that case local variations remain. The key recommendation of the article is that OAs should provide the basis of analyses of all population sub-groups unless an alternative set of zones must be used.

The values of D_{xy} and the form of the variograms reflect some distinctive features of the geographies of population sub-groups in England and Wales. For the age log-ratios, there is no evidence for strong spatial structure – this is expected as, while there are more young adults in London and other major cities and towns than in less densely populated locales, there are no spatially distinct concentrations of young or old people; the corresponding D_{xy} values are small. Ethnicity, the tenure log-ratios and CarsVans all indicate strong spatial structure while the values of D_{xy} are fairly large. In the case of Ethnicity, the main geographical feature of England and Wales relates to the locations of immigrant settlement areas (see Catney and Simpson 2010). Both housing tenure and car and van access demonstrate strong urban–rural contrasts with proportionately more social rented households in urban areas and lower rates of car and van access in urban areas (London and northern cities) and in the valleys of the former South Wales coalfields ('Welsh valleys') than elsewhere. There tend to be higher rates of no qualifications in the North and West than in the South and East, but the variogram does not indicate strong localised trends. Similarly, LLTI does not exhibit strong spatial structure, although there is strong evidence for large scale trends by poor health with generally low rates in the South East (with the exception of London) and high rates in the Welsh valleys and in urban areas of Northern England (Dorling and Thomas 2004). At local scales, however, there is little evidence for distinct spatial concentrations of people lacking qualifications or of poor health, and this is reflected in the variograms. The variograms of Employ do suggest regional trends in unemployment and this reflects generally higher rates of unemployment in the North and West than in South and East (with London as an exception). In terms of the NS-SeC log-ratios, there is evidence for regional trends (particularly in NSSEC1) and this links to the observations of Dorling and Thomas (2004) on the geography of occupational classifications with, for example, high rates of professionals and managers in London and other parts of the South East.

The analysis using the index of dissimilarity, D_{xy}, suggests that use of zones larger than OAs would be inadvisable if the concern is to assess the amount of variation between areas for age, tenure or LLTI. In contrast, larger zones could possibly be used in the analysis of (No) CarsVans and NSSEC where for each no more than 12.03% of the variation is captured by OAs. For White/BAME, only a small proportion of the variation is accounted for by the two smallest sets of zones (OA = 2.7%; LSOA = 2.03%) and this suggests that MSOAs could be used in an analysis of this simple two-part classification of ethnicity. However, there are still likely to be considerable local variations where even the use of LSOAs would result in information loss. The variogram analysis indicates that, at OA level, variables which have strong spatial structure (where the nugget effect in the fitted variogram model comprises only a small proportion of the total sill) include Ethnicity and Tenure2 while weaker spatial structure is represented by Age1, Age2, Tenure1 and LLTI. In short, this indicates that neighbouring OAs tend to be similar in terms of the proportion of White/BAME persons (Ethnicity) and of owner-occupied households as against private rented households (Tenure2) (see Section 3.2 for definitions of the log-ratios). In contrast, neighbouring OAs tend to have dissimilar proportions of people as expressed by the ratio of younger people to

those aged 65 plus (Age1), those in the two youngest age groups relative to those aged 30–64 (Age2), owner occupied and private rented households as against social rented households (Tenure1) and persons with and without a LLTI. The sill values of the variograms are reflected in the D_{xy} values (Table 3) and they show that there is considerable variation in the White/BAME population and in housing tenure. With respect to the White/BAME population, there are large differences between regions but not, on average, between local 'neighbourhoods' (i.e. semivariances are small at small lags and large at large lags). In contrast, there is much less inter-regional variation in the population by age and LLTI.

The present study shows how measured variability (represented by D_{xy} and the variogram nugget effects and structured components) and the spatial structure (represented by the variogram) differs when the zones are changed. The analysis of aggregated data is a function of the spatial structure of the data and its relation to the zonal geography (see Lloyd 2014). In a study assessing the impact in changes to the zonal system on measures of residential segregation, Wong (1997) demonstrated that changes in analytical results were a function of spatial autocorrelation in the variable and Wong (2009) illustrated this principle using synthetic data. The main focus in the study by Wong (1997) was on D_{xy}. Where there is negative spatial autocorrelation (neighbouring values tend to be dissimilar), using different zonal systems may result in quite different values of D_{xy}. In contrast, in the case of positive spatial autocorrelation, changing the forms of the zones may have little impact where the zones are smaller than the areas over which variables are positively spatially autocorrelated (see Shuttleworth *et al*. 2011). The present study identifies cases which support this assertion.

For the White/BAME variable there is considerable continuity across scales in terms of, for example, D_{xy}, and this is due to the high degree of spatial dependence (positive spatial autocorrelation) identified in the variogram analyses. It is shown that for all variables OAs, LSOAs and MSOAs display similar spatial structures at a regional scale (within 80 km), but that, for most variables (and particularly those with weaker spatial structure), a large proportion of the variation is lost with aggregation from OAs to LSOAs and from LSOAs to MSOAs. Collectively, the results provide useful information to guide future design of population surveys and also provide guidance to users in selecting an appropriate zonal system for analysis of particular variables. For example, if users want to identify distinct areas by, say, LLTI then OAs would be most appropriate. With respect to the White/BAME population, results are likely to be more robust (less sensitive to changes in zone size) and zones larger than OAs could be used with minimal loss of information (albeit with variation in information loss between areas).

In cases where there is a considerable reduction in the variation with an increase in zone size, this may suggest that geographical differences are much smaller than they are in 'reality'. As an example, using LSOAs or MSOAs in an analysis of LLTI would suggest that geographical inequalities are considerably smaller than is indicated by an analysis based on OAs. The results suggest that it is essential that zone size selection is based on the underlying characteristics of the population sub-group(s) of interest. Use of zones which are too large may provide outputs which are misleading or potentially useless with considerable economic and societal implications in terms of, for example, resource allocation and the well-being of individuals in some neighbourhoods.

In summary, using D_{xy} and the sills of the variograms, OAs are recommended for the analysis of age, housing tenure and LLTI, and, therefore, it is recommended that counts

continue to be provided for OAs. Also, for EA(Un)Emp and (No)Qual, OAs are advisable. For (No)CarsVans and NSSEC, OAs are not essential for univariate analysis and LSOAs, or even MSOAs would be appropriate for many analyses. Finally, for White/BAME, even LAs would capture most (83%) of the variation although a 'loss' of 17% of the variation is still considerable; as such, MSOAs would be more suitable for univariate analysis. Any of OAs, LSOAs and MSOAs can be used to represent the overall geography of each variable (the variogram ranges are similar within each log-ratio for OAs, LSOAs and MSOAs), but the choice of zone is important if the magnitude of differences between areas is a concern, as it is with analyses of geographical inequalities. The results suggest that OAs (constructed based on tenure of household and dwelling type) do a good job of spatially partitioning areas in that, for tenure, much of the variation is captured at OA level. OAs should be used wherever possible in analyses of all population sub-groups and users should be aware that even using LSOAs may result in considerable loss of information and thus flawed findings. Where there is a prior reason to prefer aggregations of OAs, these should be selected and utilised with care and with consideration of the sub-groups included in the analyses and the ways in which they are spatially distributed across the study area.

The analyses presented here were based on broad groups; it is clear that the implications of changing the zonal system will be even greater for some smaller groups (e.g. specific housing tenures or ethnic groups) than they are for the aggregations employed here. Future analysis will seek to build on this analysis by using smaller groupings. The research could also be expanded to consider the impact of changing zones on multivariate analyses. Another issue which would benefit from attention would be the loss of information associated with using administrative zones such as wards; using wards seems likely to result in greater information loss than using LSOAs or MSOAs as wards are an administrative geography and not, unlike OAs (and LSOAs and MSOAs), constructed based on population characteristics.

Information about appropriate spatial scales is important, particularly in the light of debates about the future of the UK Census. The analyses presented here provide a means of assessing how much information might be lost if alternatives to the most recent Census model in the United Kingdom (i.e. 2011) provide less rich spatial detail. Of course, the Census may not provide the best source of information on population characteristics (e.g. see Norman and Bambra 2007 for a discussion about the use of sickness benefit data as a regularly updatable indicator of health over small areas). While this article takes as its focus the Census, there is no reason that the approaches assessed could not be applied in assessing alternative means of capturing information on the population of the United Kingdom over small areas to help ascertain how these approaches may produce outputs which can be used to properly characterise the geographies of population sub-groups and address important issues about geographic inequalities and allocation of resources to those in need.

Notes

1. Bulletin for CCGs: Issue 31, 28 March 2013: http://www.england.nhs.uk/2013/03/28/ccg-bulletin-issue-31/
2. http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/output-area–oas-/index.html

3. http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/super-output-areas--soas-/index.html
4. With the caveats that the age range for qualifications differs between 2001 (16–74) and 2011 (16+) and NS-SeC counts for 2011 include imputed responses and are not directly comparable with those for 2001.
5. For BAME, nearly 50% of the action is accounted for by LAs while for NoCarsVans the figure is 33%. For all other counts the figure is below 25%. For most counts more than 40% of the action is accounted for by OAs with the largest figure for NoLLTI (64%).

Acknowledgements

The Office for National Statistics is thanked for provision of the data on which the analyses were based. Office for National Statistics, 2011 Census: Aggregate data (England and Wales) [computer file]. UK Data Service Census Support. Downloaded from: http://infuse.mimas.ac.uk. This information is licensed under the terms of the Open Government Licence [http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2]. Digitised Boundary Data (England and Wales) [computer file]. UK Data Service Census Support. Downloaded from: http://edina.ac.uk/census. The comments of the anonymous reviewers are acknowledged gratefully.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

Part of the research on which this article was based was supported by the Economic and Social Research Council [grant ES/L014769/1] and this is acknowledged gratefully.

References

Aitchison, J., 1986. *The statistical analysis of compositional data*. London: Chapman & Hall.

Atkinson, P.M. and Curran, P.J., 1997. Choosing an appropriate spatial resolution for remote sensing investigations. *Photogrammetric Engineering and Remote Sensing*, 63, 1345–1351.

Atkinson, P.M. and Tate, N.J., 2000. Spatial scale problems and geostatistical solutions: a review. *The Professional Geographer*, 52, 607–623. doi:10.1111/0033-0124.00250

Catney, G., 2015. Exploring a decade of small area ethnic (de-)segregation in England and Wales. *Urban Studies* doi:10.1177/0042098015576855

Catney, G. and Simpson, L., 2010. Settlement area migration in England and Wales: assessing evidence for a social gradient. *Transactions of the Institute of British Geographers*, 35, 571–584. doi:10.1111/tran.2010.35.issue-4

Dorling, D. and Rees, P., 2003. A nation still dividing: the British Census and social polarisation 1971–2001. *Environment and Planning A*, 35, 1287–1313. doi:10.1068/a3692

Dorling, D. and Thomas, B., 2004. *People and places: a 2001 Census atlas of the UK*. Bristol: The Policy Press.

Duncan, O.D. and Duncan, B., 1955. A methodological analysis of segregation indexes. *American Sociological Review*, 20, 210–217. doi:10.2307/2088328

Egozcue, J.J. and Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37, 795–828. doi:10.1007/s11004-005-7381-9

Egozcue, J.J. and Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. *In*: A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, eds. *Compositional data analysis in the geosciences: from theory to practice*. Geological Society Special Publications No 264. London: Geological Society, 145–160.

Egozcue, J.J., *et al*., 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300. doi:10.1023/A:1023818214614

Folch, D.C. and Spielman, S.E., 2014. Identifying regions based on flexible user-defined constraints. *International Journal of Geographical Information Science*, 28, 164–184. doi:10.1080/13658816.2013.848986

Fotheringham, A.S., Brunsdon, C., and Charlton, M., 2002. *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester: John Wiley and Sons.

Goovaerts, P., 2008. Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Mathematical Geosciences*, 40, 101–128. doi:10.1007/s11004-007-9129-1

Goovaerts, P., Jacquez, G.M., and Greiling, D., 2005. Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms. *Geographical Analysis*, 37, 152–182. doi:10.1111/gean.2005.37.issue-2

Grasland, C., Mathian, H., and Vincent, J.-M., 2000. Multiscalar analysis and map generalisation of discrete social phenomena: statistical problems and political consequences. *Statistical Journal of the United Nations Economic Commission for Europe*, 17, 157–188.

Griffith, D.A., Wong, D.W.S., and Whitfield, T., 2003. Exploring relationships between the global and regional measures of spatial autocorrelation. *Journal of Regional Science*, 43, 683–710. doi:10.1111/j.0022-4146.2003.00316.x

Lloyd, C.D., 2012. Analysing the spatial scale of population concentrations by religion in Northern Ireland using global and local variograms. *International Journal of Geographical Information Science*, 26, 57–73. doi:10.1080/13658816.2011.563743

Lloyd, C.D., 2014. *Exploring spatial scale in geography*. Chichester: John Wiley and Sons.

Lloyd, C.D., 2015. Assessing the spatial structure of population variables in England and Wales. *Transactions of the Institute of British Geographers*, 40, 28–43. doi:10.1111/tran.2014.40.issue-1

Lloyd, C.D., Pawlowsky-Glahn, V., and Egozcue, J.J., 2012. Compositional data analysis in population studies. *Annals of the Association of American Geographers*, 102, 1251–1266. doi:10.1080/00045608.2011.652855

Lloyd, C.D., Shuttleworth, I.G., and Wong, D.W.S., eds., 2014. *Social-spatial segregation: concepts, processes and outcomes*. Bristol: Policy Press.

Manley, D., 2014. Scale, aggregation, and the modifiable areal unit problem. *In*: M.M. Fischer and P. Nijkamp, eds. *Handbook of regional science*. Berlin: Springer-Verlag, 1157–1171.

Manley, D., Flowerdew, R., and Steel, D., 2006. Scales, levels and processes: studying spatial patterns of British Census variables. *Computers, Environment and Urban Systems*, 30, 143–160. doi:10.1016/j.compenvurbsys.2005.08.005

Martin, D., 1989. Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers, New Series*, 14, 90–97.

Martin, D., Nolan, A., and Tranmer, M., 2001. The application of zone-design methodology in the 2001 UK Census. *Environment and Planning A*, 33, 1949–1962. doi:10.1068/a3497

Massey, D.S. and Denton, N.A., 1988. The dimensions of residential segregation. *Social Forces*, 67, 281–315. doi:10.1093/sf/67.2.281

Mockus, A., 1998. Estimating dependencies from spatial averages. *Journal of Computational and Graphical Statistics*, 7, 501–513.

Moellering, H. and Tobler, W., 1972. Geographical variances. *Geographical Analysis*, 4, 34–50. doi:10.1111/j.1538-4632.1972.tb00455.x

Nagle, N.N., Sweeney, S.H., and Kyriakidis, P.C., 2011. A geostatistical linear regression model for small area data. *Geographical Analysis*, 43, 38–60. doi:10.1111/j.1538-4632.2010.00807.x

NHS England, 2014. *Technical guide to the formulae for 2014-15 and 2015-16 revenue allocations to clinical commissioning groups and area teams*. Leeds: NHS England Strategic Finance Available from: http://www.england.nhs.uk/wp-content/uploads/2014/03/tech-guide-rev-allocs.pdf [Accessed 17 August 2015].

Norman, P., 2010. Demographic and deprivation change in the UK, 1991–2001. *In*: J. Stillwell, *et al*., eds. *Spatial and social disparities*. Understanding Population Trends and Processes, Vol. 2. Dordrecht: Springer, 17–35.

Norman, P. and Bambra, C., 2007. Incapacity or unemployment? The utility of an administrative data source as an updatable indicator of population health. *Population, Space and Place*, 13, 333–352. doi:10.1002/(ISSN)1544-8452

ONS (Office for National Statistics), 2013a. *The Census and future provision of population statistics in England and Wales: public consultation*. Available from: http://www.ons.gov.uk/ons/about-ons/get-involved/consultations-and-user-surveys/consultations/beyond-2011-consultation/consultation-document–english-version.pdf [Accessed 16 January 2014].

ONS (Office for National Statistics), 2013b. *Beyond 2011: narrowing down the options*. Available from: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011–narrowing-down-the-options–o3-.pdf [Accessed 10 August 2015].

ONS (Office for National Statistics), 2014. *Beyond 2011: final options report* (O4). Available from: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011–final-options-report.pdf [Accessed 10 August 2015].

Openshaw, S., 1984. *The modifiable areal unit problem. Concepts and techniques in modern geography 38*. Norwich: GeoBooks.

Openshaw, S. and Taylor, P.J., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *In*: N. Wrigley, ed. *Statistical applications in the spatial sciences*. London: Pion, 127–144.

Pawlowsky, V. and Burger, H., 1992. Spatial structure analysis of regionalized compositions. *Mathematical Geology*, 24, 675–691. doi:10.1007/BF00894233

Pawlowsky-Glahn, V. and Olea, R.A., 2004. *Geostatistical analysis of compositional data*. New York: Oxford University Press.

Pebesma, E.J. and Wesseling, C.G., 1998. Gstat: a program for geostatistical modelling, prediction and simulation. *Computers & Geosciences*, 24, 17–31. doi:10.1016/S0098-3004(97)00082-4

Reardon, S.F., *et al*., 2009. Race and space in the 1990s: changes in the geographic scale of racial residential segregation, 1990–2000. *Social Science Research*, 38, 55–70. doi:10.1016/j.ssresearch.2008.10.002

Reardon, S.F., *et al*., 2008. The geographic scale of metropolitan racial segregation. *Demography*, 45, 489–514. doi:10.1353/dem.0.0019

Shuttleworth, I.G., Lloyd, C.D., and Martin, D.J., 2011. Exploring the implications of changing census output geographies for the measurement of residential segregation: the example of Northern Ireland 1991–2001. *Journal of the Royal Statistical Society: Series A*, 174, 1–16. doi:10.1111/j.1467-985X.2010.00647.x

Silk, J., 1981. *The analysis of variance. Concepts and techniques in modern geography 30*. Norwich: GeoAbstracts.

Spielman, S.E., Folch, D., and Nagle, N., 2014. Patterns and causes of uncertainty in the American Community Survey. *Applied Geography*, 46, 147–157. doi:10.1016/j.apgeog.2013.11.002

Tranmer, M. and Steel, D., 2001. Using local census data to investigate scale effects. *In*: N.J. Tate and P.M. Atkinson, eds. *Modelling scale in geographical information science*. Chichester: John Wiley and Sons, 105–122.

Voas, D. and Williamson, P., 2000. The scale of dissimilarity: concepts, measurement and an application to socio-economic variation across England and Wales. *Transactions of the Institute of British Geographers, New Series*, 25, 465–481. doi:10.1111/tran.2000.25.issue-4

Webster, R. and Oliver, M.A., 2007. *Geostatistics for environmental scientists*. 2nd ed. Chichester: John Wiley and Sons.

Wong, D., 2009. The modifiable areal unit problem (MAUP). *In*: A.S. Fotheringham and P.A. Rogerson, eds. *The SAGE handbook of spatial analysis*. London: SAGE Publications, 105–123.

Wong, D.W.S., 1997. Spatial dependency of segregation indices. *The Canadian Geographer*, 41, 128–136. doi:10.1111/j.1541-0064.1997.tb01153.x

Zhang, J., Atkinson, P.M., and Goodchild, M.F., 2014. *Scale in spatial information and analysis*. Boca Raton: CRC Press.